

**گزارش تمرین شماره 1**

**درس داده کاوی**

**جناب آقای دکتر فراهانی**

**جناب آقای دکتر خرد پیشه**

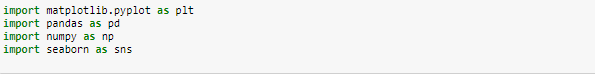
**نسرین هداوند سوری**

**19/12/1399**

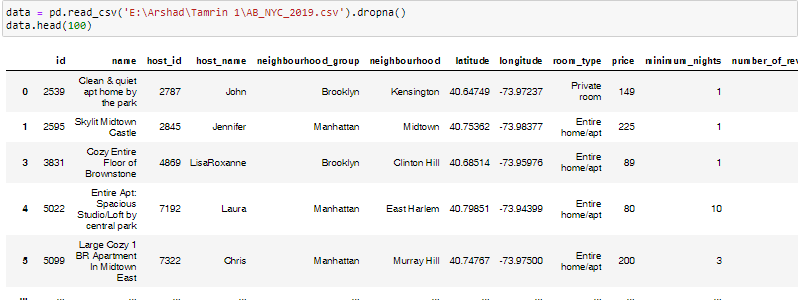
بررسی مجموعه داده اول :

توصیف : مجموعه داده اول مربوط به داده های جمع آوری شده از خانه ای اجاره ای برای اقامت کوتاه مدت در شهر نیویورک آمریکا است و در آن اطلاعاتی در مورد میزبان ها، میهمان ها، مکان اقامت گاه، زمان و مدت اجاره ، قیمت اجاره و ... وجود دارد.

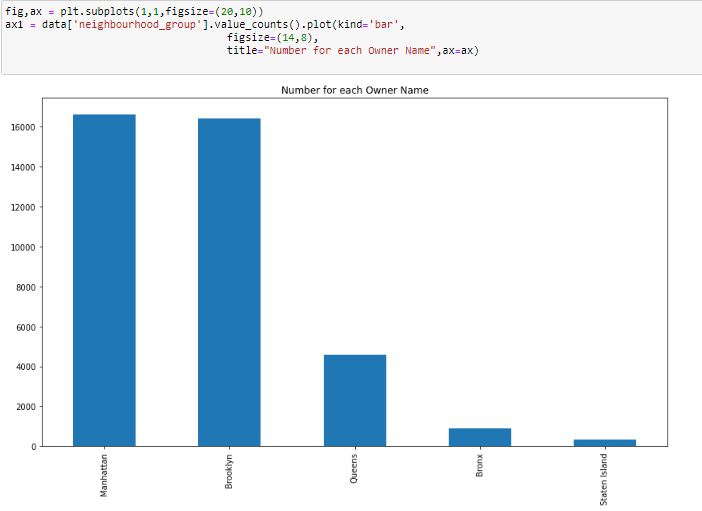
بررسی سوال 1 » در مورد میزبانها و مناطق مختلف و ارتباط آنها و بررسی ویژگی های آن :

ابتدا پکیج های مورد نیاز فراخوانی کردم :

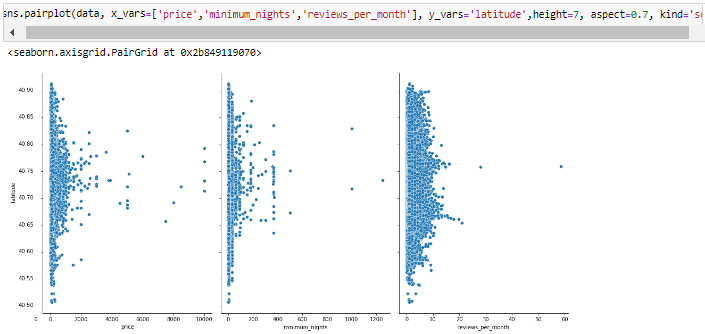
سپس مجموعه داده اول فراخوانی کردم :



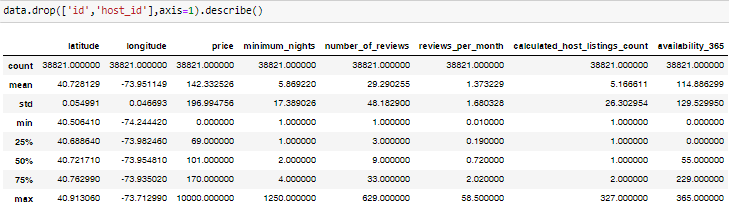
سپس "neighbourhood\_group" بررسی کردم که تو لیست که تعداد ثبت شده از میزبان ها در نواحی مختلف نیویورک به چه صورت است که در نمودار قابل مشاهده است در مجموعه داده حدود 38000 sample وجود داشت که 32000 تا مربوط به manhatan و Brooklyn است و نزدیک 4000 مربوط به Queens و کمتر از 2000 مربوط به Bronx و Staten Island



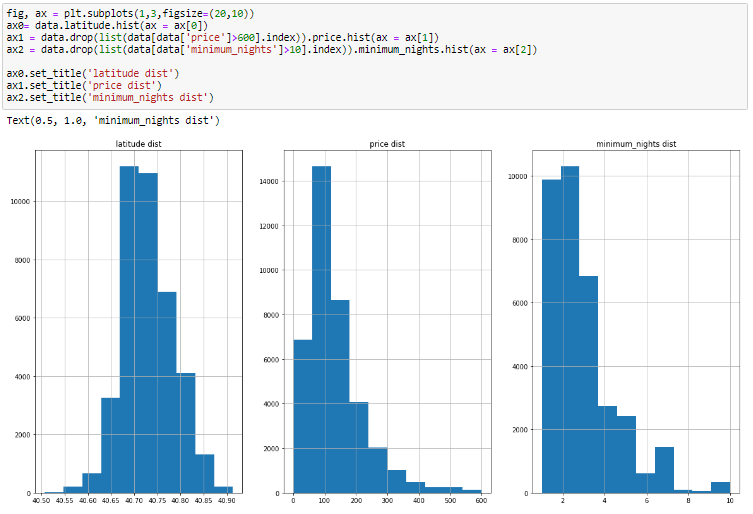
سپس سوال یک به صورت دیگری بر اساس عرض جغرافیایی بررسی کردم این که price چه ارتباطی با عرض جغرافیایی دارد یا minimum\_night چه ارتباطی با عرض جغرافیایی دارد و Reviews\_per\_month با عرض جغرافیایی دارد که در نمودار مشخص است که در نواحی میانی عرض جغرافیایی تقریبا بیشتر شده است و می توان گفت نواحی میانه قسمت پرطرفدار در شهر نیویورک است .

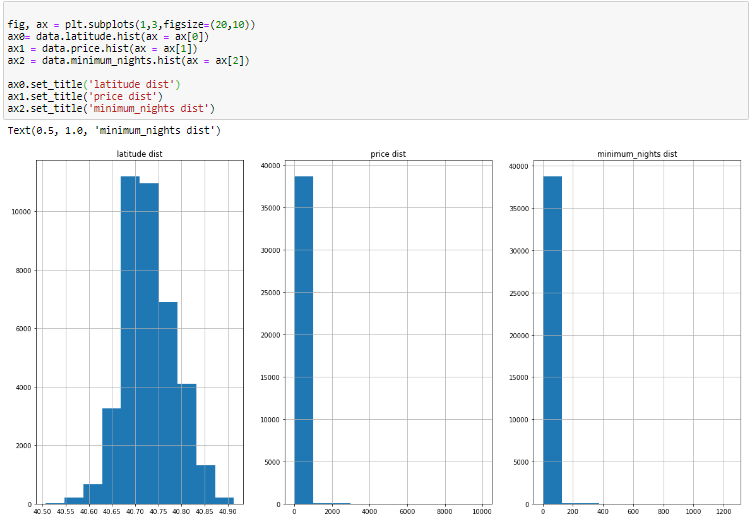


بررسی سوال 2 » بررسی مشخصات آماری : مجموع تمام داده ها و میانگین و تابع توزیع و مینیمم و بررسی 25 و 50 و 75 درصد داده و ماکسیمم



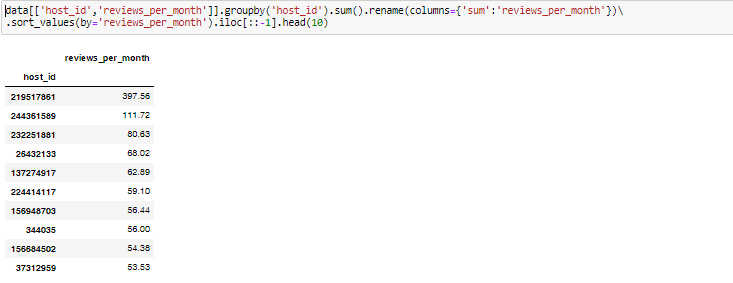
برای سوال 2 سه تا ویژگی انتخاب کردم و تابع توزیع رسم کردم

 ابتدا بازه ها خوب نبود و بازه ها را کوچکتر کردم تا پراکندگی زیاد و احتمال وقوع تعداد بیشتر می شود .



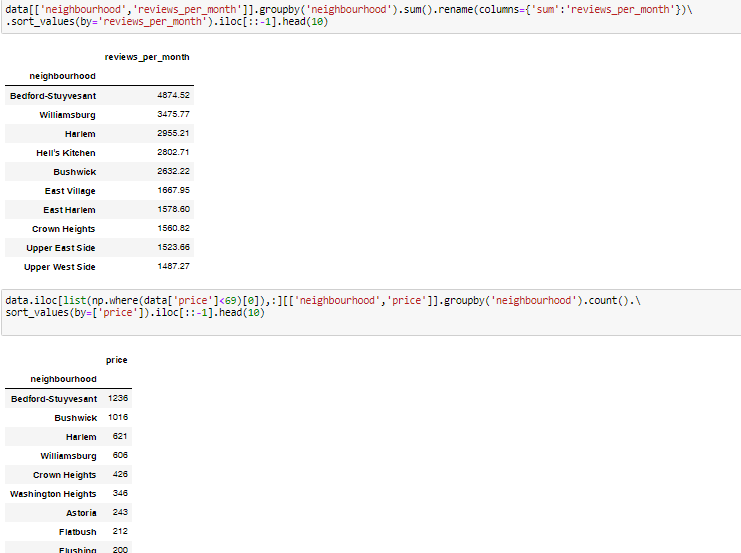
بررسی سوال 3 » کدوم host شلوغ تر است و چرا؟

من بر اساس host\_id بررسی کردم و به نظرم اونی که review\_per\_month از همه بیشتر شلوغ تر بوده



بررسی سوال 4 » آیا تفاوت قابل توجهی در ترافیک در مناطق مختلف وجود دارد و چه چیزی می تواند دلیل آن باشد؟

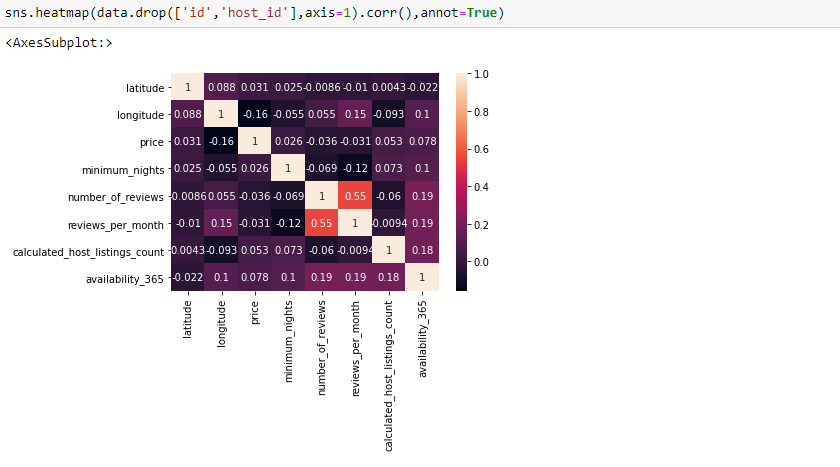
برای پاسخ این سوال اون محله هایی از شهر نیویورک به همراه review\_per\_month بررسی کردم و به نظرم جاهایی که review\_per\_month بیشتر است احتمال وقوع ترافیک بیشتر است و بعد بر اساس price بررسی کردم و بر اساس کم ترین price آن محله هایی که ارزان تر است اسکان در آن زیاد هست و ترافیک هم بیشتر هست و به همین دلیل است .



طرح سوال 5 »

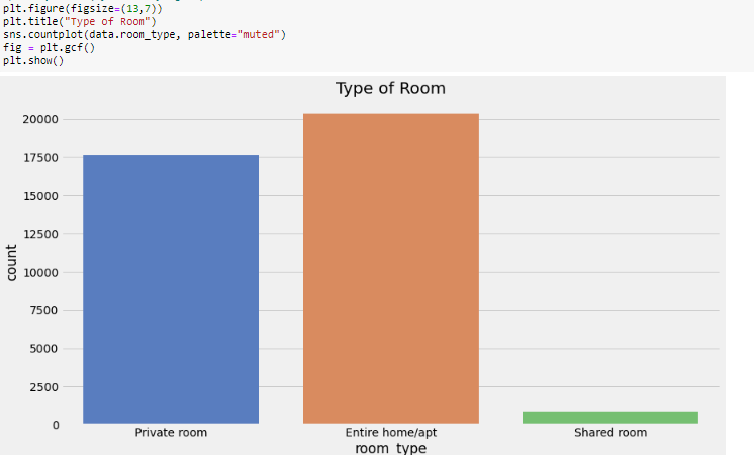
کدام عوامل اثر بیشتری داردکه یک مسافر تعداد شب های بیش تری را اقامت داشته باشد؟

ارتباط بین ویژگی ها و اینکه نسبت به هم به چه شکل تغییر می کند.



طرح سوال 6 »

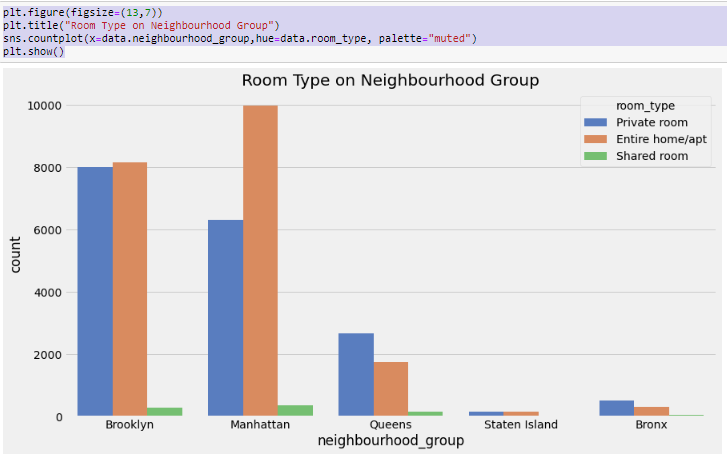
می توانیم ببینیم که خانه / آپارتمان بیشترین سهم را دارد و پس از آن اتاق خصوصی است و کمترین ترجیح مربوط به اتاق مشترک است.



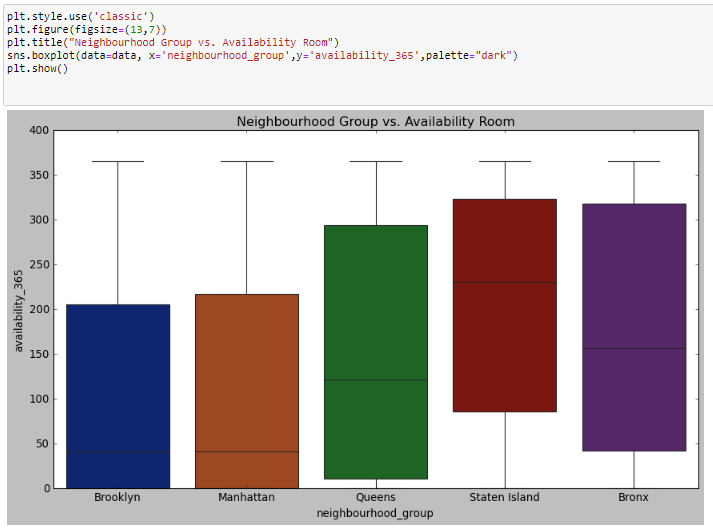
طرح سوال 7 »

بررسی نوع اقامت در هر محله :

نمودار نشان می دهد که خانه / آپارتمان در منهتن بیشتر از همه است ، در حالی که اتاق های خصوصی و آپارتمان های نزدیک بروکلین تقریباً برابر هستند.



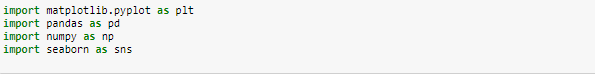
طرح سوال 8»  
بررسی ارتباط بین availability room and neighborhood group و می توان گفت در staten Island داده ها مدین بالایی دارد پس در بیشتر روزهای سال خانه های این بخش در دسترس است و بر عکس در بروکلین تعداد کمی در بازه 200 تا 365 روز قابل دسترس هستند.



بررسی مجموعه داده دوم :

توصیف : این مجموعه داده مربوط به مسابقات فوتبال بین المللی است که در مقابل مسابقات جام جهانی، جام های قاره ای، تورنومنت ها، بازی های دوستانه و ... انجام شده است.به این منظوراطلاعات مختلفی از قبیل، نام تیم ها، محل انجام مسابقه، میزبان و میهمان ، زمان بازی، تعداد گل ها ، نتیجه بازی و ... ذخیره شده اند.

بررسی سوال 1 »

ابتدا پکیج های مورد نیاز فراخوانی کردم :

سپس مجموعه داده اول فراخوانی کردم :



بررسی سوال 1 »

بهترین تیم در تمام دوران ها ؟

بهترین تیم تیمی است که بیشترین برد را داشته باشد .تیم برزیل بیشترین برد را داشته است.

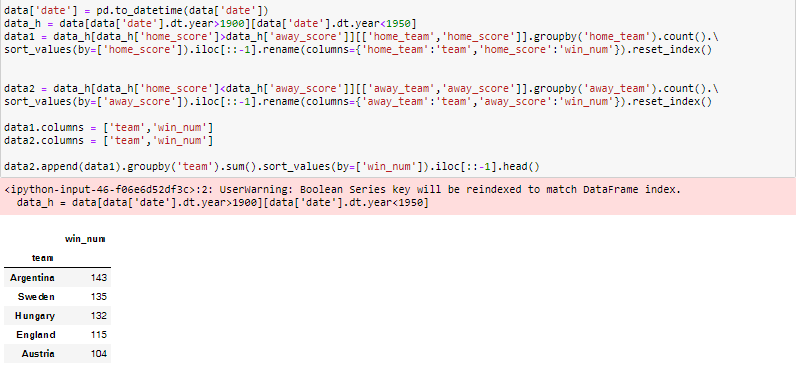


بررسی سوال 2 »

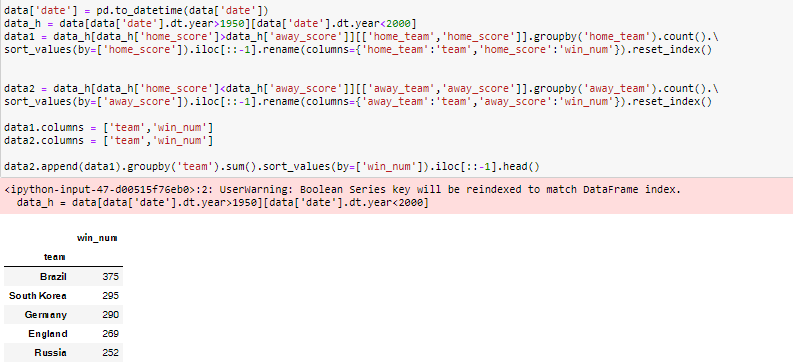
کدام تیم ها در دوره های مختلف فوتبال سلطه داشتند.

در چند بازه زمانی بررسی کردم یکی کمتر از سال 1990

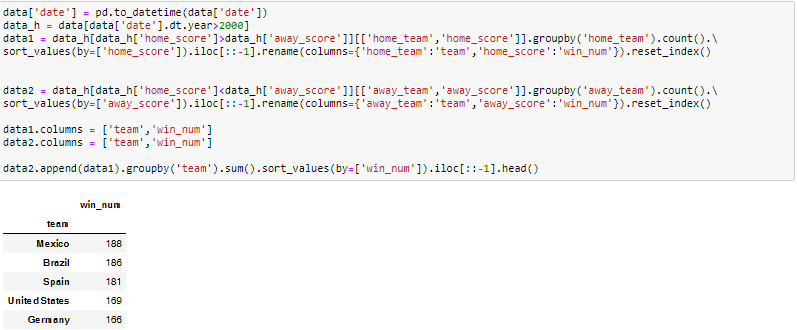
بعد بازه زمانی بزرگتر از 1990 و کمتر از 1950



بعد بازه زمانی از کمتر از 1950 و بیشتر از 2000



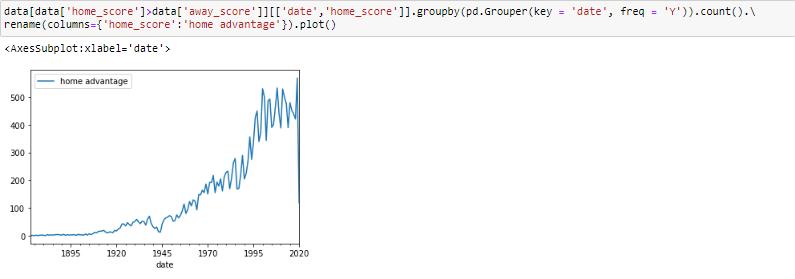
بعد بازه زمانی بزرگتر از سال 2000



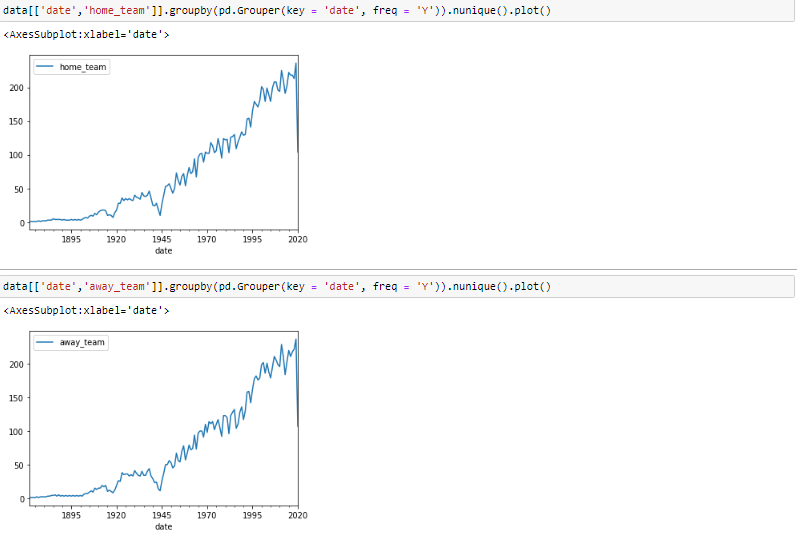
بررسی سوال 3 »

چه روندهایی در فوتبال بین المللی در طول سالها وجود داشته است، بازی خانگی ، کل گل زده ، توزیع قدرت تیم ها و غیره

برای این به این شکل بررسی کردم که بازی خانگی چقدر تاثیر روی برد داشته. که می بینیم در واقع در طی سالها و گذشتن سال ها اثر بهتری داشته تا تیم های میزبان بازی را ببرن .



بعد در ادامه اومدم بررسی کردم تعداد تیم های مختلف که بازی می کنن چطوری تغییر کردن که می بینیم که هرچی زمان بیشتری گذشته تعداد کشورهای بیشتری اومدن فوتبال انجام دادن.



بررسی سوال 4 »

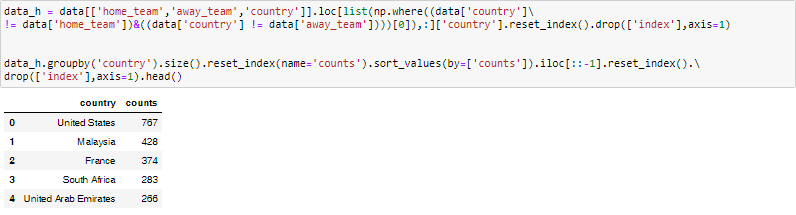
آیا می توانیم چیزی در مورد ژئوپلیتیک از مسابقات فوتبال بگوییم-چگونه تعداد کشورها تغییر کرده است ،کدام تیم ها دوست دارند با یکدیگر بازی کنند؟

من برداشتم از این سوال این هست که کدوم تیم ها بیشتر مایل بودن با هم بازی کنن و من تیم هایی که بیشترین بازی را با هم داشتن را بررسی کردم و از خروجی آن به نظرم کشورهایی که همسایه هستن بیشتر با هم بازی کرده اند.



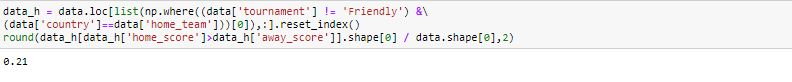
بررسی سوال 5 »

کدام کشورهای بازی های بیشتری را میزبانی کردند که خودشان در آن بازی نکرده اند.



بررسی سوال 6 »

تو تورنومنت های غیر دوستانه بازی های رسمی تیم هایی که میزبان بودن چقدر شانس بردن بوده است .



بررسی سوال 7 »

کدام تیم ها بیش ترین بازی دوستانه داشتند و اثرش کمک کننده یا مخرب :

تیم ها را برحسب تعداد تیم هایی که بازی دوستانه داشتن مرتب کردم که در خروجی خوب تیم هایی که تو فوتبال مطرح هستن نمایش داده شده که به همین دلیل کمک کننده بوده است .



طرح سوال 8 »

تو بازی های مختلف بین تیم های مختلف تعداد گل های مختلف تیم میزبان با دقت بیش تری می توان پیش بینی کرد یا تیم مهمان ؟

در این سوال می توان گفت که در یک مجموعه داده هرچی واریانس کم تر باشد و پراکندگی داده ها کم تر و سپس احتمال و دقت پیش بینی بالاتر می رود و هر چی واریانس بیشتر پراکندگی داده بیش تر و دقت پیش بینی کم تر می شود و می بینیم تیم میزبان انحراف معیار کم تر است

